基于推荐技术的图书管理系统的设计

周强[1]　许丽媛[1]　孔贝贝[1]　李曦[2]
（1.　中国科学院文献情报中心　北京 100190）
（2.　中国科学技术大学　安徽　230026）

[摘要]

目前的图书管理系统，没有把图书检索和图书推荐结合起来，本文介绍的图书管理系统，对图书进行了分类，提供了按书名、作者进行检索，在检索结果页面中，显示推荐结果。本文介绍了该图书管理系统的设计，该网站的推荐系统采用了基于物品的协同过滤算法、基于内容的推荐算法，本文详细介绍了这两个算法，接着，本文全面介绍了该图书管理系统中推荐系统部分的设计。现在，该图书管理系统已经完成了概要设计和详细设计。

[关键词]

基于物品的协同过滤算法　图书　推荐　基于内容的推荐算法

Based on the recommendation of the design of books management system

Zhou Qiang[1]　Xu LiYuan[1]　Kong BeiBei[1]　Li Xi[2]
(1. National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
(2. University of Science and Technology of China, Anhui 230026, China)

[Abstract]

The current library management system, did not combine searching book with recommending book , library management system described in this paper, the books were classified, and provides a search by title, author, and from the search results page and displays the results. This article introduced the design of books management system management system, the site recommendation system based on collaborative filtering, content-based recommendation algorithm, these algorithms are described in detail in this paper, then, this article provides an overview of the recommended system design of the library management system. Now that the library management system has completed the preliminary design and detailed design.

[Keyword]

collaborative filtering algorithm based on items
book　recommendation　Content-based recommendation algorithm

一、图书管理系统的介绍

书籍是人类进步的阶梯，书籍已成为传播知识、科学技术和保存文化的重要工具。

由于图书的非常多，本文设计的图书管理系统，收录的图书包括电子技术、

计算机科学技术、图书情报等领域的图书，收录的图书有相对经典的图书，如经典的教材名著，还有最新技术的图书，如云计算、移动互联网等。本文提出的设计方案，把图书检索和图书推荐结合起来，是实验性的方案，用来论证图书检索和图书推荐结合的效果。

二、现有的图书管理系统简述

2.1、北京大学图书馆检索系统

北京大学图书馆的馆藏宏大丰富、学科齐全、珍品荟萃。到 2011 年底，文献资源累积量约 1,100 余万册（件），其中纸质藏书 800 余万册，以及近年来大量引进和自建的国内外数字资源，包括各类数据库、电子期刊、电子图书和多媒体资源约 300 余万册（件）。

检索结果中，图书的详情页面包括图书封面、书名、作者、出版日期、索书号、总页数、馆藏所在地、借阅状况。

图书的详情页面没有推荐信息。

2.2、清华大学图书馆检索系统

清华大学图书馆的馆藏，到 2013 年底，总量约有 463.0 万册（件），形成了
以自然科学和工程技术科学文献为主体，兼有人文、社会科学及管理科学文献等多种类型、多种载体的综合性馆藏体系。除中外文图书外，馆藏资源还包括：古籍线装书、期刊、本校博士硕士论文、缩微资料等。

检索结果有检索列表，检索列表页面有图书封面，可以预约；图书详情页面，包括作者、书名、出版社、出版日期、摘要、目录、馆藏所在地、借阅状况、图书封面。

图书的详情页面没有推荐信息。

2.3、当当网

当当网（www.dangdang.com）是全球知名的综合性网上购物商城。图书是当当网的主营业务之一，在库图书、音像商品超过 80 万种。目前当当网的注册用户遍及全国 32 个省、市、自治区和直辖市。

检索结果有检索列表，检索列表页面有图书封面；图书详情页面，包括图书封面、书名、作者、出版社、出版日期、ISBN、页数、摘要、作者简介、目录、在线试读。

图书的详情页面提供了两种推荐数据："看过本商品的还看了"、"买过本商品的还买了"。

三、图书管理系统设计

图书管理系统，包括了搜索、推荐系统，本文重点介绍推荐系统[1-2,5-9]，推荐

算法采用基于物品的协同过滤算法。

数据存储在关系数据库中，本文采用 MySql 数据库，为了加快检索速度，创建了 Lucene 索引，检索时，从 Lucene 索引中读取数据。

Web Server 采用 Tomcat。

在系统的最前端，配置了 Squid，来进行反向代理，通过反向代理来进行负载均衡，部署了两套系统，包括数据库、Lucene 索引、web server，数据库 、Lucene 索引、web server 这两套系统是一致的。数据库的操作是追加，可以实时进行，Lucene 索引、web server 代码的更新是轮流在凌晨更新的。

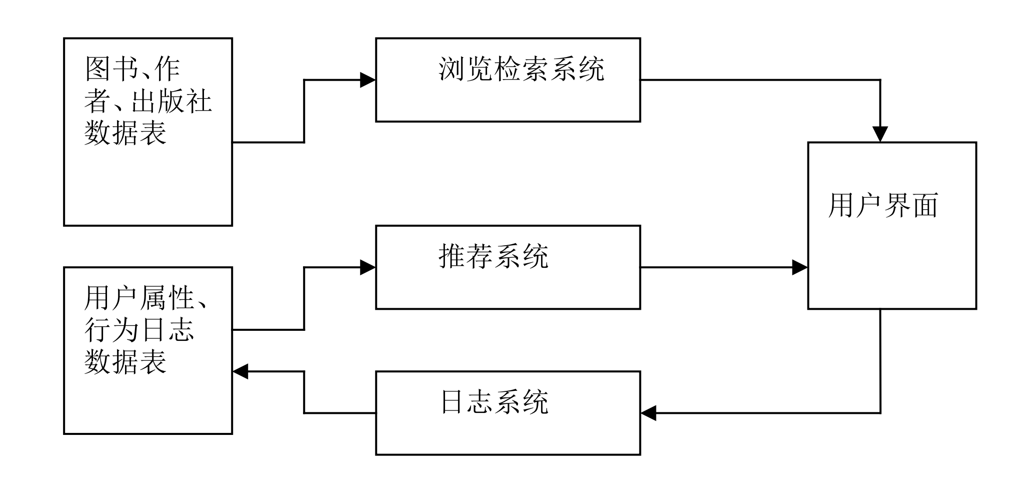


图 1、图书管理系统

图书数据表存储着图书的数据，包括书名、作者、出版社、出版日期、ISBN、页数、摘要、作者简介、前言、目录。图书的封面文件部署在文件系统中。

作者数据表存储着作者名字、作者简介。

出版社数据表存储着出版社名字、出版社简介。

书名、作者、出版社数据表提供数据给浏览检索系统，为了加快检索速度，创建了索引。

日志系统对于推荐系统是很重要的。

用户通过用户界面进行浏览检索时，把用户的行为写入日志了；用户通过用户界面借阅时，把用户的行为也写入日志了。

本文的推荐系统是根据用户的日志来计算并返回推荐结果，推荐结果展示是图书的详细信息界面的“浏览本书资料的还浏览了”、“借阅本书的还借阅了”两个模块部分。

四、推荐算法简述

图书的数目非常多，从大量图书中找到自己感兴趣的图书是一件相对困难的事情。推荐系统的任务是联系用户和信息，一方面帮助用户发现对自己有价值的信息，另一方面让信息能够展现在对它感兴趣的用户面前。

本系统采用的推荐算法[1-13]是基于物品的协同过滤算法和基于内容的过滤算

法。基于内容的过滤算法采用基于向量的表示方法，即余弦定理。

算法的步骤如下：先用基于物品的协同过滤算法计算出物品列表；接着用余弦定理对这个物品列表再进行排序。

4.1、基于物品的协同过滤算法

本系统中的物品，指的是图书。

基于物品（图书）的协同过滤算法主要分为两步：

1、计算图书之间的相关度；

2、根据图书的相关度和用户的历史行为给用户生成推荐列表。

物品相似度的公式：

$$w_{ij} = |N(i)| \cap |N(j)| \sqrt{|N(i)||N(j)|}$$

其中 $N(i)$ 是喜欢图书 i 的用户数，$N(j)$ 是喜欢图书 j 的用户数。

计算物品相似度的步骤：

1)、建立用户-图书倒排表（即每个用户建立一个他喜欢的图书的列表）；

2)、对于每个用户，把他图书列表中的图书两两在共现矩阵 C 中加 1，其中 $C[i][j]$记录了同时喜欢图书 i 和图书 j 的用户数。

3)、把矩阵 C 归一化，得到图书之间的余弦相似度矩阵 W。

在得到图书之间的相似度后，通过如下公式计算用户 u 对一个图书的兴趣：

$$P_{uj} = \sum_{i \in N(u) \cap S(j,K)} w_{ji} r_{ui}$$

这里 $N(u)$ 是用户喜欢的图书的集合。$S(i,K)$ 是和图书 i 最相似的 K 个图书的集合，w_{ji} 是图书 j 和 i 的相似度，r_{ui} 是用户 u 对图书 i 的兴趣。

这些计算是在离线的情况下计算的。

这样，可以构造图书—图书的倒排索引。

4.2、余弦定理

余弦定理，需要预先把词汇表导入到数据库中，设这个词汇表的总数为 m。

主要分为四步：

1、对于图书摘要和前言中的所有实词，计算它们的 TF/IDF（单文本词汇频率/逆文本频率值）；

2、按照这些实词在词汇表的位置对它们的 TF/IDF 值排序；

3、如果词汇表中的某个词在图书摘要和前言中没有出现，对应的值为零那么这词汇表的总数 m，组成一个 m 维的向量。我们就用这个向量来代表这本图书，并成为图书的特征向量。如果两本图书的特征向量相近，则对应的图书内容相似，它们在推荐列表的排名位置就接近。

4、用余弦定理，来计算向量的相似度，设图书 X 和图书 Y 的对应向量分

别是 $x_1, x_2, ..., x_m$ 和 $y_1, y_2, ..., y_m$，向量夹角的余弦等于，

$$\cos\theta = (x_1 y_1 + x_2 y_2 + ... + x_m y_m)/(\sqrt{x_1^2 + x_2^2 + ... + x_m^2}\sqrt{y_1^2 + y_2^2 + ... + y_m^2})$$

当两本图书向量夹角的余弦等于 1 时，这两本图书完全重复；当夹角的余弦接近于一时，两本图书相似，从而在推荐列表的排名位置就接近；夹角的余弦越小，两本图书在推荐列表的排名位置就越远。

五、推荐系统设计

下面是推荐系统[14-22]的架构图，说明了数据的流向。

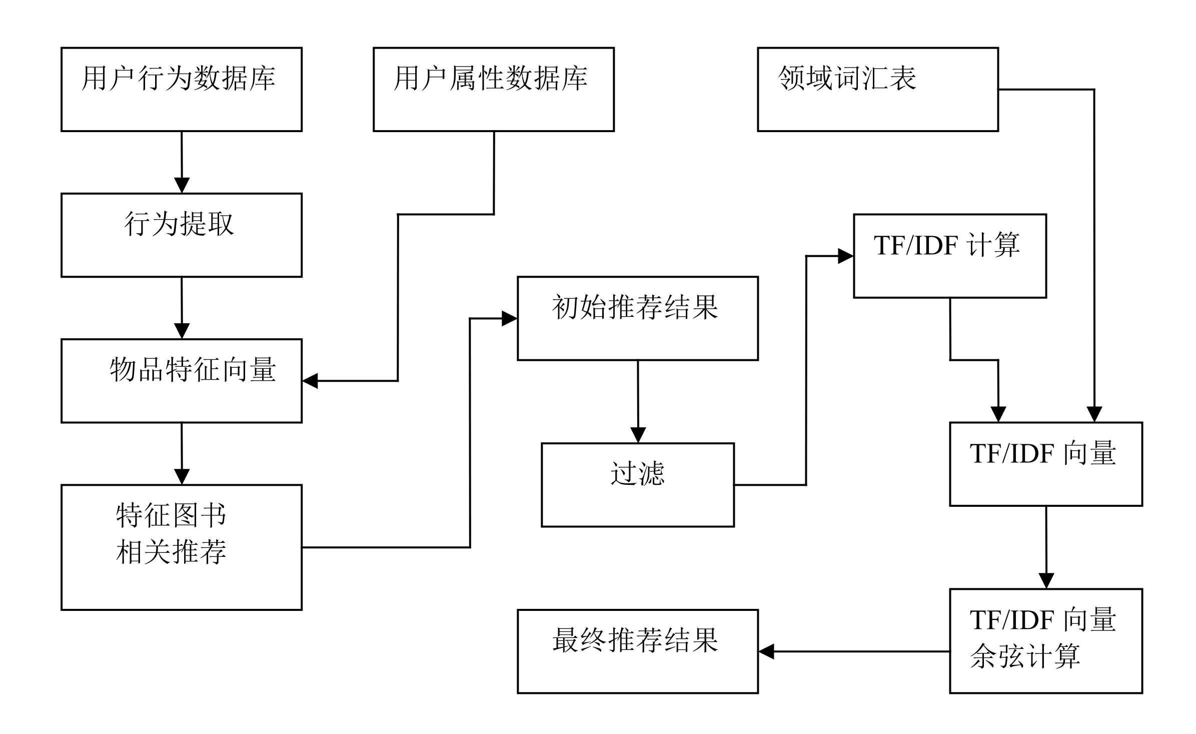


图 2、推荐系统设计

用户的特征包括两种，一种是从用户的注册信息中提取出来的，即用户的人口统计学特征；另一种特征主要是从用户的行为中计算出来。

一个物品特征向量由特征以及特征的权重组成，在计算时需要考虑以下因素。

1、 用户行为分为浏览图书、借阅图书两种，其中借阅图书的权重大。
2、 用户行为产生的时间，用户近期的浏览、借阅行为比较重要。
3、 用户行为的次数，用户会浏览一本图书很多次，用户对同一本图书的浏览的次数反映了用户对图书的兴趣。浏览次数多的图书对应的特征权重高，一本图书的浏览人数越多，则该书对应的特征权重就越高；对于同

一本书，借阅图书的特征权重比浏览图书的特征权重高。

4、 图书的热门程度，如果用户对一个热门图书产生了行为，有可能是跟风，可能对该图书没有太大的兴趣，因此，对于不热门图书的权重高。

在得到用户的特征向量后，我们可以根据离线的相关表得到初始的图书推荐列表，其存储格式如下所示：

特征标识、图书标识、书名、作者、权重。

在得到初步的推荐列表后，需要过滤掉不符合要求的图书—质量不好的图书。

经过过滤后的推荐结果，采用基于内容的推荐算法进行下一步推荐。

对于图书摘要和前言中的所有实词，计算它们的 TF/IDF （单文本词汇频率/逆文本频率值），按照这些实词在词汇表的位置对它们的 TF/IDF 值排序，生成图书的特征向量，接着用余弦定理计算向量相似度。

最后生成最终推荐列表，这推荐列表时离线生成的，为了加快查询速度，把这些推荐列表的数据写入 Lucene 索引中。

六、总结

本文首先介绍了图书管理系统的需求，接着简述了现有的图书管理系统。在这些的基础上，介绍了图书管理系统的设计，并说明了基于物品的协同过滤算法和基于内容的过滤算法，接着说明了推荐系统的设计方案。

[参考文献]

[1] Marco Degemmis,Pasquale Lops,Giovanni Semeraro. A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation[J]. User Modeling and User - Adapted Interaction . 2007，3：217-255.

[2] Rosario Girardi,Leandro Balby Marinho. A domain model of Web recommender systems based on usage mining and collaborative filtering[J]. Requirements Engineering . 2007，1：23-40.

[3] Jiawei Han,Jian Pei,Yiwen Yin,Runying Mao. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach[J]. Data Mining and Knowledge Discovery . 2004，1：53-87.

[4] Ken Goldberg,Theresa Roeder,Dhruv Gupta,Chris Perkins. Eigentaste: A Constant Time Collaborative Filtering Algorithm[J]. Information Retrieval . 2001 ，2：133-151.

[5]马宏伟,张光卫,李鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统,2009,07:1282-1288.

[6]王嫣然,陈梅,王翰虎,张鑫. 一种基于内容过滤的科技文献推荐算法[J]. 计算机技术与发展,2011,02:66-69.

[7]刘枚莲,刘同存,李小龙. 基于用户兴趣特征提取的推荐算法研究[J]. 计算机应用研究,2011,05:1664-1667.

[8]杨博,赵鹏飞. 推荐算法综述[J]. 山西大学学报(自然科学版),2011,03:337-350.

[9]邢哲,梁竞帆,朱青. 多维度自适应的协同过滤推荐算法[J]. 小型微型计算机系统,2011,11:2210-2216.

[10]刘旭东,陈德人,王惠敏. 一种改进的协同过滤推荐算法[J]. 武汉理工大学学报(信息与管理工程版),2010,04:550-553.

[11]赵亮,胡乃静,张守志. 个性化推荐算法设计[J]. 计算机研究与发展,2002,08:986-991.
[12]郭艳红,邓贵仕. 协同过滤的一种个性化推荐算法研究[J]. 计算机应用研究,2008,01:39-41+58.
[13]李聪,梁昌勇,马丽. 基于领域最近邻的协同过滤推荐算法[J]. 计算机研究与发展,2008,09:1532-1538.
[14] 王国霞、刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用，2012，7：66-76.
[15] 许海玲、吴潇、李晓东、阎保平. 互联网推荐系统比较研究[J]. 软件学报，2009，2：350-362.
[16] 刘建国、周涛,汪秉宏.个性化推荐系统的研究进展[J].自然科学进展,2009,01:1-15.
[17] 刘鲁、任晓丽. 推荐系统研究进展及展望[J]. 信息系统学报，2008，1：82-90.
[18] 陈定权、朱维凤. 关联规则与图书馆书目推荐[J]. 情报理论与实践，2009，6：81-84.
[19] 李树青、徐侠、许敏佳. 基于读者借阅二分网络的图书可推荐质量测度方法及个性化图书推荐服务[J]. 中国图书馆学报，2013，3：83-95.
[20] 范旭. 以豆瓣网和中国国家图书馆为案例的网上书目推荐系统研究[J]. 图书馆学研究，2008，8：44-48.
[21] 丁雪. 基于数据挖掘的图书智能推荐系统研究[J]. 情报理论与实践，2010，5：107-110.
[22] 王义、马尚才. 基于用户行为的个性化推荐系统的设计与应用[J]. 计算机系统应用，2010，8：29-33.